Persistent Target Tracking Using Likelihood Fusion in Wide-Area and Full Motion Video Sequences

Rengarajan Pelapur,* Sema Candemir,* Filiz Bunyak,* Mahdieh Poostchi,*
Guna Seetharaman,† Kannappan Palaniappan*
Email: {rvpnc4, mpr69}@mail.missouri.edu, {candemirs, bunyak, palaniappank}@missouri.edu
Gunasekaran.Seetharaman@rl.af.mil
*Dept. of Computer Science, University of Missouri, Columbia, MO 65211, USA
†Air Force Research Laboratory, Rome, NY 13441, USA

***Abstract*—Vehicle tracking using airborne wide-area motion imagery (WAMI) for monitoring urban environments is very challenging for current state-of-the-art tracking algorithms, compared to object tracking in full motion video (FMV). Characteristics that constrain performance in WAMI to relatively short tracks range from the limitations of the camera sensor array including low frame rate and georegistration inaccuracies, to small target support size, presence of numerous shadows and occlusions from buildings, continuously changing vantage point of the platform, presence of distractors and clutter among other confounding factors. We describe our Likelihood of Features Tracking (LoFT) system that is based on fusing multiple sources of information about the target and its environment akin to a track-before-detect approach. LoFT uses image-based feature likelihood maps derived from a template-based target model, object and motion saliency, track prediction and management, combined with a novel adaptive appearance target update model. Quantitative measures of performance are presented using a set of manually marked objects in both WAMI, namely Columbus Large Image Format (CLIF), and several standard FMV sequences. Comparison with a number of single object tracking systems shows that LoFT outperforms other visual trackers, including state-of-the-art sparse representation and learning based methods, by a significant amount on the CLIF sequences and is competitive on FMV sequences.**

I. Introduction

Target tracking remains a challenging problem in computer vision [1] due to target-environment appearance variabilities, significant illumination changes and partial occlusions. Tracking in aerial imagery is generally harder than traditional tracking due to the problems associated with a moving platform including gimbal-based stabilization errors, relative motion where sensor and target are both moving, seams in mosaics where stitching is inaccurate, georegistration errors, and drift in the intrinsic and extrinsic camera parameters at high altitudes [2]. Tracking in Wide-Area Motion Imagery (WAMI) poses a number of additional difficulties for vision-based tracking algorithms due to very large gigapixel sized images, low frame rate sampling, low resolution targets, limited target contrast, foreground distractors, background clutter, shadows, static and dynamic parallax occlusions, platform motion, registration, mosacing across multiple cameras, object dynamics, etc. [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. These difficulties make the tracking task in WAMI more challenging compared to standard ground-based or even narrow field-of-view (aerial) full motion video (FMV).

Traditional visual trackers either use motion/change detection or template matching. Persistent tracking using motion detection-based schemes need to accommodate dynamic behaviors where initially moving objects can become stationary for short or extended time periods, then start to move again. Motion-based methods face difficulties with registration, scenes with dense set of objects or near-stationary targets. Accuracy of background subtraction and track association dictate the success of these tracking methods [10], [9], [14], [15]. Template trackers on the other hand, can drift off target and attach themselves to objects that seem similar, without an update to the appearance model [16], [2].

Visual tracking is an active research area with a recent focus on appearance adaptation, learning and sparse representation. Appearance models are used in [17], [18], [19], [20], classification and learning techniques have been studied in [21], [22], and parts-based deformable templates in [23]. Gu *et al.* [20] stress low computation cost in addition to robustness and propose a simple yet powerful Nearest Neighbor (NN) method for real-time tracking. Online multiple instance learning (MILTrack) is used to achieve robustness to image distortions and occlusions [21]. The P-N tracker [22] uses bootstrapping binary classifiers and shows higher reliability by generating longer tracks. Mei *et al.* [24], [25] propose a robust tracking method using a sparse representation approach within a particle filter framework to account for pose changes.

We have developed the Likelihood of Features Tracking (LoFT) system to track objects in WAMI. The overall LoFT tracking system shown in Figure 1, can be broadly organized into several categories including: (i) Target modeling, (ii) Likelihood fusion, and (iii) Track management. Given a target of interest, it is modeled using a rich feature set including intensity/color, edge, shape and texture information [4], [26]. The novelty of the overall LoFT system stems from a combination of critical components including a flexible set of features to model the target, an explicit appearance update scheme, adaptive posterior likelihood fusion for track-before-detect, a kinematic motion model, and track termination working cooperatively in balance to produce a reliable tracking system.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302 Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Persistent Target Tracking Using Likelihood Fusion in Wide-Area and Full Motion Video Sequences	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory,Rome,NY,13441**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 15th International Conference on Information Fusion held in Sinapore on 9-12 July 2012. Sponsored in part by Office of Naval Research and Office of Naval Research Global.

14. ABSTRACT
Vehicle tracking using airborne wide-area motion imagery (WAMI) for monitoring urban environments is very challenging for current state-of-the-art tracking algorithms, compared to object tracking in full motion video (FMV). Characteristics that constrain performance in WAMI to relatively short tracks range from the limitations of the camera sensor array including low frame rate and georegistration inaccuracies, to small target support size, presence of numerous shadows and occlusions from buildings, continuously changing vantage point of the platform, presence of distractors and clutter among other confounding factors. We describe our Likelihood of Features Tracking (LoFT) system that is based on fusing multiple sources of information about the target and its environment akin to a track-before-detect approach. LoFT uses image-based feature likelihood maps derived from a template-based target model object and motion saliency, track prediction and management combined with a novel adaptive appearance target update model. Quantitative measures of performance are presented using a set of manually marked objects in both WAMI, namely Columbus Large Image Format (CLIF), and several standard FMV sequences. Comparison with a number of single object tracking systems shows that LoFT outperforms other visual trackers including state-of-the-art sparse representation and learning based methods, by a significant amount on the CLIF sequences and is competitive on FMV sequences.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

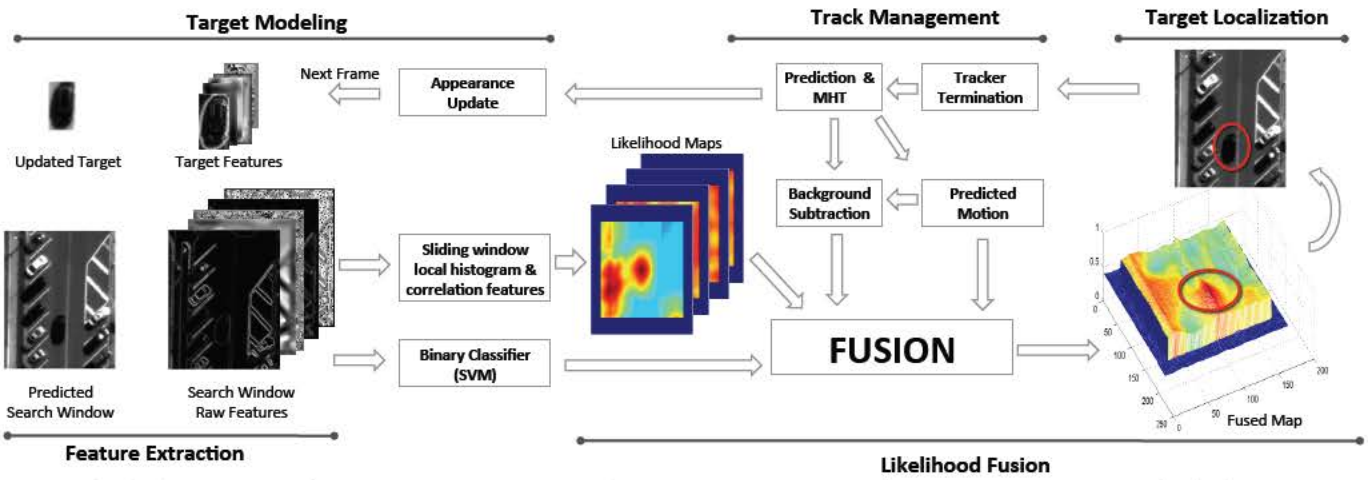

Fig. 1. Likelihood of Features Tracking (LoFT) processing pipeline showing major components including feature extraction, feature likelihood map estimation by combining with the template, vehicle detection using support vector machine (SVM) classification, fusion module that also incorporates prediction based motion and background subtraction based motion, to produce a fused likelihood for target localization after track extension. The track management includes termination module, prediction with or without multiple hypothesis tracking (MHT) and object appearance updating for adaptive target modeling.

The rest of the paper is organized as follows. Section II describes estimating features and fusing the posterior likelihood maps. Section III describes the novel target appearance modeling and adaptive update modules. Section IV describes the tracker management component that is often lacking in other systems including smooth trajectory assessment and appropriate tracker termination to maintain track purity. Experimental results for both CLIF WAMI and FMV video sequences are described in Section V followed by conclusions.

II. Likelihood Fusion

The target area is modeled using features that can be grouped into categories such as block, edge, shape and texture based. We use block-based intensity features, gradient magnitude for edge-based features, gradient orientation information using Histogram of Oriented Gradients, eigenvalues of the Hessian matrix for shape information, and local binary patterns for texture. See [4] for more details on these feature descriptors and their integral histogram-based implementation.

LoFT uses a recognition-based target localization approach using the maximum likelihood of the target being within the search region conditioned on a feature. A likelihood map is estimated for each feature by comparing feature histograms of the target within the search region using a sliding window-based approach (see Fig. 1). Each pixel in the likelihood map indicates the posterior probability of that pixel belonging to the target. Fusing features enables adaptation of the tracker to dynamic environment changes and target appearance variabilities. Using a track-before-detect approach provides more robust localization especially for cluttered dense environments [27]. Feature adaptation uses a weighted sum Bayes fusion rule that tends to perform better than other methods such as the product rule [28]. The critical aspect in weighted sum fusion is the relative importance of feature maps. Each feature performs differently depending on the target characteristic and environmental situations during tracking. Equally weighted fusion of likelihood maps can decrease performance, when some of the features are not informative in that environment. The importance assigned to each feature can be adapted to the changes in target pose and the surrounding background. Temporal feature weight adaptation can improve performance under changes that are not explicitly modeled by the tracker.

We considered two weighting schemes including the Variance Ratio (VR) [19] and the Distractor Index [4]. LoFT fuses the histogram-based and correlation-based features in two stages. First, histogram-based features are fused using the VR method [19], [29] which adaptively weights the features according to the discriminative power between the target and the background measured using the two-class ratio of total to within class variances. Second, non-histogram (*i.e.* correlation) based features are combined with the fused histogram-based features using the Distractor Index method proposed by Palaniappan *et al.* [4]. In this method, the number of local maxima within 90% of the peak likelihood and within the spatial support of the object template, $\mathcal{N}_T$, are used as the number of viable peaks for the i^{th} feature, $m_i \in [1, \infty)$. Fusion feature weights in LoFT are then calculated using [4],

$$w_i \approx m_i^{-1} (\sum_{i=1}^{n} 1/m_i)^{-1}. \quad (1)$$

Consequently, high distractor index values will result in low weights for unreliable features. By assuming the environment does not change drastically across frames, the system fuses the likelihood maps of frame k using the feature weights which were estimated at frame $k-1$. Calculating feature weights dynamically enables the tracker to cope with small appearance changes in target and environment. Strong local maxima in the fused map which exceeds a predetermined threshold are considered as potential target locations.

III. Target Modeling

LoFT [4] uses the principle of single target template-based tracking where target features are used to match an area or region in subsequent frames. Static template-based tracking has been studied in computer vision since the 1970's. Currently, generative models such as [17], [18] or discriminative

models such as [30], [21] all have online and offline versions to robustly adapt to object appearance variability. Recently, several trackers based on sparse representation have shown promise in handling complex appearance changes [25], [31], [32]. Our dynamic appearance adaptation scheme maintains and updates a single template by estimating affine changes in the target to handle orientation and scale changes [33], using multiscale Laplacian of Gaussian edge detection followed by segmentation to largely correct for drift. Multi-template extensions of the proposed approach are straightforward but computationally more expensive. LoFT is being extended in this direction by parallelization of the integral histogram [34].

A. Appearance Update

Given a target object template, T_s, in the initial starting image frame, I_s, we want to identify the location of the template in each frame of the WAMI sequence using a likelihood matching function, $M(\cdot)$. Once the presence or absence of the target has been determined, we then need to decide whether or not to update the template. The target template needs to be updated at appropriate time points during tracking, without drifting off the target, using an update schedule which is a tradeoff between plasticity (fast template updates) and stability (slow template updates). The template search and update model can be represented as,

$$\mathbf{x}^*_{k+1} = \arg\max_{\mathbf{x}\in\mathcal{N}_W} M(I_{(k+1,c\in\mathcal{N}_T)}(\mathbf{x}+c), T_u),\ k \geq s,\ \ u \geq s \tag{2}$$

$$T_{k+1} = \begin{cases} I_{(k+1,}(\mathbf{x}^*_{k+1}+c), & \text{if } \ f(\mathbf{x}^*_{k+1}, I_{k+1}, T_u)) > Th \\ \mathbf{c}\in\mathcal{N}_T) & \\ T_u, & \text{otherwise} \end{cases}$$

where $M(\cdot)$ denotes the posterior likelihood estimation operator that compares the vehicle/car template from time step u, T_u (with support region or image chip, $c \in \mathcal{N}_T$), within the image search window region, $\mathcal{N}_W$, at time step $k+1$. The optimal target location in I_{k+1} is given by $\mathbf{x}^*_{k+1}$. If the car appearance is stable with respect to the last updated template, T_u, then no template update for, T_{k+1}, is performed. However, if the appearance change function, $f(\cdot)$, is above a threshold indicating that the object appearance is changing and we are confident that this change is not due to an occlusion or shadow, then the template is updated to the image block centered at $\mathbf{x}^*_{k+1}$. Instead of maintaining and updating a single template model of the target a collection of templates can be kept (as in learning-based methods) using the same framework, in which case we would search for the best match among all templates in Eq. 2. Note that if $u = s$ then the object template is never updated and remains identical to the initialized target model; $u = k$ naively updates on every frame. Our adaptive update function $f(\cdot)$ considers a variety of factors such as orientation, illumination, scale change and update method.

In most video object tracking scenarios the no update scheme rarely leads to better performance [17] whereas naively updating on every frame will quickly cause the tracker to drift especially in complex video such as WAMI [4]; making the tradeoff between these two extremes is commonly referred to as the stability-plasticity dilemma [35]. Figure 2 shows several frames of a sample car from the CLIF sequences as its appearance changes over time. Our approach to this dilemma is to explicitly model appearance variation by estimating scale and orientation changes in the target that is robust to illumination variation. Segmentation can further improve performance [36], [37].

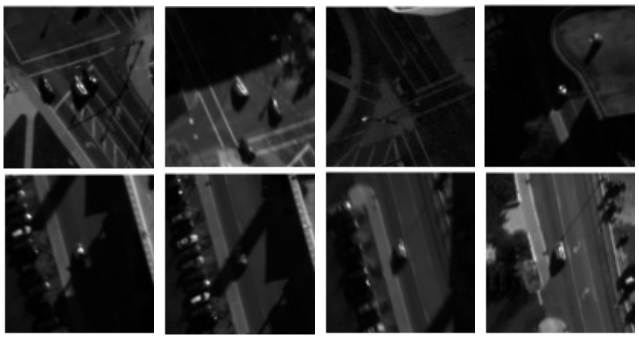

Fig. 2. Orientation and intensity appearance changes of the same vehicle over a short period of time necessitates updates to the target template at an appropriate schedule balancing plasticity and stability.

We recover the affine transformation matrix to model the appearance update by first extracting a reliable contour of the object to be tracked using a multiscale Laplacian of Gaussian, followed by estimating the updated pose of the object using the Radon transform projections as described below.

B. Laplacian of Gaussian

We use a multi-scale Laplacian of Gaussian (LoG) filter to increase the response of the edge pixels. Using a series of convolutions with scale-normalized LoG kernels $\sigma^2\nabla^2 G(x, y, \sigma^2)$ where σ denotes the standard deviation of the Gaussian filter,

$$I_{k,L}(x, y, \sigma^2) = I_k(x, y) * \sigma^2\nabla^2 G(x, y, \sigma^2) \tag{3}$$

we estimate the object scale at time k by estimating the mean of the local maxima responses in the LoG within the vehicle template region $\mathcal{N}_T$. If this $\hat{\sigma}^*_k$ has changed from $\hat{\sigma}^*_u$ then the object scale is updated.

C. Orientation estimation

The Radon transform is used to estimate the orientation of the object [33] and applying the transform on the LoG image $I_{k,L}(x, y)$, we can denote the line integrals as:

$$R_k(\rho, \theta) = \iint I_{k,L}(x, y)\delta(\rho - x\cos\theta - y\sin\theta)\ dx\ dy \tag{4}$$

where $\delta(\cdot)$ is the Dirac delta function that samples the image along a ray (ρ, θ). Given the image projection at angle θ, we estimate the variance of each projection profile and search for the maximum in the projection variances by using a second-order derivative operator to achieve robustness to illumination change [38]. An example of vehicle orientation and change in orientation estimation is shown in Figure 3. This appearance update procedure seems to provide a balance between plasticity and stability that works well for vehicles in aerial

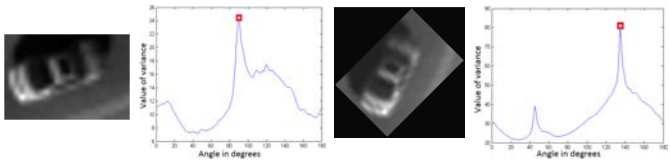

Fig. 3. Vehicle orientations are measured wrt vertical axis pointed up. (a) Car template. (b) Variance of Radon transform profiles with maximum at 90° (red sq). (c) Car template rotated by 45° CCW. (d) Peak in variance of Radon transform profiles at 135° (red sq), for correct change in orientation of 45°.

imagery. More detailed performance evaluation of orientation estimation is found in our related work [39].

IV. Track Management

A robust tracker should maintain track history information and terminate the tracker as performance deteriorates irrecoverably (e.g. camera seam boundary), the target leaves the field-of-view (e.g. target exiting the scene), enters a long occluded/shadow region, or the tracker has lost the target. LoFT incorporates multiple *track termination* conditions to ensure high precision (track purity) and enable downstream tracklet stitching algorithms to operate efficiently during *track stitching*. Track linearity or smoothness guides the tracker to select more plausible target locations incorporating vehicle motion dynamics and a module for terminating the tracker.

A. Smooth Trajectory Dynamics Assumption

Peaks in the fused likelihood map are often many due to clutter and denote possible target locations including distractors. However, only a small subset of these will satisfy the smooth motion assumption (*i. e.* linear motion). Checks for smooth motion/linearity is enforced before a candidate target location is selected to eliminate improbable locations. Figure 4 illustrates the linear motion constraint. The red point indicates a candidate object with a very similar appearance to the target being tracked, but this location is improbable since it does not satisfy the trajectory motion dynamics check and so the next highest peak is selected (yellow dot). This condition enforces smoothness of the trajectory thus eliminating erratic jumps and does not affect turning cars.

Fig. 4. When the maximum peak (red dot) deviates from the smooth trajectory assumption (in this case linearity) LoFT ignores the distractor to select a less dominant peak satisfying the linearity constraint (yellow dot).

B. Prediction & Filtering Dynamical Model

LoFT can use multiple types of filters for motion prediction. In the implementation evaluation for this paper we used a Kalman filter for smoothing and prediction [40], [41] to determine the search window in the next frame, I_{k+1}. The Kalman filter is a recursive filter that estimates the state, $\mathbf{x}_k$, of a linear dynamical system from a series of noisy measurements, $\mathbf{z}_k$. At each time step k the state transition model is applied to the state to generate the new state,

$$\mathbf{x}_{k+1} = \mathbf{F}_k\,\mathbf{x}_k + \mathbf{v}_k \tag{5}$$

assuming a linear additive Gaussian process noise model. The measurement equation under uncertainty generates the observed outputs from the true ("hidden") state.

$$\mathbf{z}_k = \mathbf{H}_k\,\mathbf{x}_k + \mathbf{w}_k \tag{6}$$

where $\mathbf{v}_k$ denotes process noise (Gaussian with zero-mean and covariance $\mathbf{Q}_k$), $\mathbf{w}_k$ denotes measurement noise (Gaussian with zero-mean and covariance $\mathbf{R}_k$). The system plant is modeled by known linear systems, where $\mathbf{F}_k$ is the state-transition matrix and $\mathbf{H}_k$ is the observation model.

Possible target locations within the search window are denoted by peak locations in the fused posterior vehicle likelihood map. Candidate locations are then filtered by incorporating the prediction information. Given a case where feature fusion indicates low probability of the target location (due to occlusions, image distortions, inadequacy of features to localize the object, etc.) the filtering-based predicted position is then reported as the target location. Figure 5 shows LoFT with the appearance-based update module being active over the track segments in yellow with informative search windows, whereas in the shadow region the appearance-based features become unreliable and LoFT switches to using only filtering-based prediction mode (track segments in white).

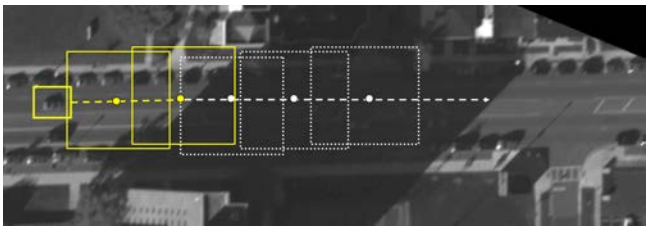

Fig. 5. Adaptation to changing environmental situations. LoFT switches between using fused feature- and filterin-based target localization (yellow boxes) within informative search windows (yellow boxes) and predominantly filtering based localization in uninformative search windows (white boxes).

C. Target vs Environment Contrast

LoFT measures the dissimilarity between the target and its surrounding environment in order to assess the presence of occlusion events. If the VR between the target and its environment is below a threshold, this indicates a high probability that the tracker/target is within an occluded region. In such situations, LoFT relies more heavily on the dynamical filter predictions. Figure 6 shows a sample frame which illustrates the difference between high and low VR locations.

D. Image/Camera Boundary Check

LoFT determines if the target is leaving the scene, crossing a seam or entering an image boundary region on every iteration in order to test for the disappearance of targets. If the

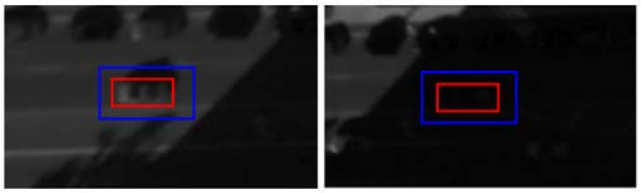

Fig. 6. Pixels within the red rectangle form the foreground (Fg) distribution, pixels between the red and blue rectangles form the background (Bg) distribution. Left: High VR when Fg and Bg regions have different distributions. Right: Low VR when Fg and Bg regions have similar distributions.

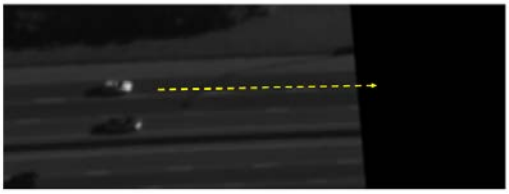

Fig. 7. Termination of tracks for targets leaving the working image boundary.

predicted location is out of the working boundary, the tracker automatically terminates to avoid data access issues (Figure 7).

V. Experimental Results

A. Datasets Used

LoFT was evaluated using the Columbus Large Image Format (CLIF) [42] WAMI dataset which has a number of challenging conditions such as shadows, occlusions, turning vehicles, low contrast and fast vehicle motion. We used the same vehicles selected in [11] which have a total of 455 ground-truth locations of which more than 22% are occluded locations. The short track lengths combined with a high degree of occlusions makes the tracking task especially challenging. Several examples of the difficulties in vehicle tracking in CLIF are illustrated in Figure 8. Figure 9 shows that half the sequences in this sample set of tracks have a significant amount of occluded regions and Table I summarizes the challenges in each sequence. We used several FMV sequences which have been used to benchmark a number of published tracking algorithms in the literature. These sequences include: 'girl', 'david', 'faceocc', 'faceocc2' [20] and allow comparison of LoFT against a number of existing tracker results for which source code may not be available.

B. Registration and Ground-Truth for CLIF WAMI

In our tests we used the same homographies as in [11] that were estimated using SIFT (Scale Invariant Feature Transform) [43] with RANSAC to map each frame in a sequence to the first base frame. Several other approaches have been used to register CLIF imagery including Lucas-Kanade, and correlation-based [44], or can be adapted for WAMI [45], [46]. Using these homographies we registered consecutive frames to the first frame in each sequence. The homographies when applied to the ground-truth bounding boxes can produce inaccurate quadrilaterals since these transformations are on a global frame level. All quadrilaterals were automatically replaced with axis aligned boxes and visually inspected to manually replace any incorrect bounding boxes, on registered frames, with accurate axis aligned boxes using KOLAM [5], [6], [47] or MIT Layer Annotation Tool [48].

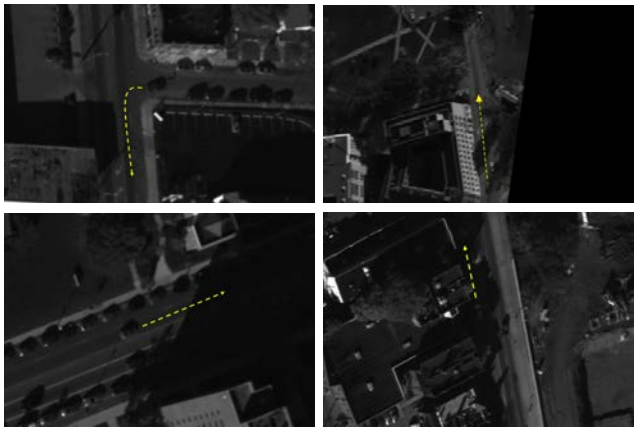

Fig. 8. Example of challenging conditions: Target appearance changes during turning (C4-1-0), low contrast and shadows (C3-3-4), shadow occlusion (C0-3-0) and combined building and shadow occlusion (C2-4-1) [49].

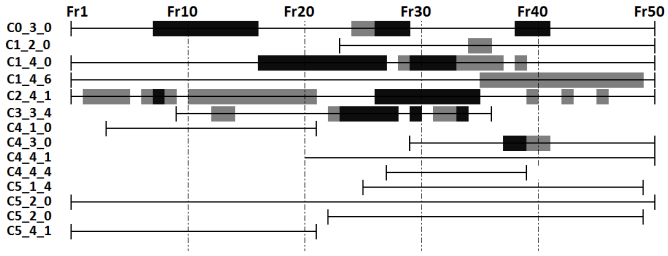

Fig. 9. Distribution of occluded frames in the 14 CLIF seq. Black: fully occluded, Gray: partially occluded. Target is occluded in 22.4% of the frames.

Seq. No	Challenges	Track Length	Target Size [pixel]	Occ.Fr
C0_3_0	Occlusion	50	17x25	17
C1_2_0	Occlusion	27	21x15	2
C1_4_0	Occlusion	50	21x17	21
C1_4_6	Occlusion	50	25x25	15
C2_4_1	Occlusion	50	25x17	32
C3_3_4	Occlusion	27	27x17	12
C4_1_0	Turning car	18	15x25	-
C4_3_0	Occlusion	20	21x17	3
C4_4_1	Low contrast	30	17x21	-
C4_4_4	-	13	17x25	-
C5_1_4	Fast target motion	23	27x11	-
C5_2_0	Fast target motion	49	21x15	-
C5_3_7	-	27	27x47	-
C5_4_1	Low Contrast	21	27x19	-
Total		455		102

TABLE I
CHARACTERISTICS OF THE 14 CLIF SEQUENCES SUMMARIZED FROM [11] SHOWING TRACK LENGTH, VEHICLE TARGET SIZE AND NUMBER OF OCCLUDED FRAMES. IMAGE FRAMES ARE 2008 × 1336 PIXELS.

C. Quantitative Comparison

We used several retrieval or *detection-based* performance metrics to evaluate the trackers. The first one is the Missing Frame Rate (MFR), which is the percentage of number of *missing* frames to the total number of ground-truth frames,

$$\text{MFR} = \frac{\#\ missing\ frames}{\#\ total\ GT\ frames} \tag{7}$$

A frame is labeled as *missing*, if the detected/estimated object location with associated bounding box overlaps with the ground-truth bounding box by less than 1% or there is no estimated bounding box at all (*i.e.* due to early track termination). The one percent overlap threshold is the correct one that was actually used in the CLIF experiments reported in Ling *et al.* [11] not 50% (personal communication). We used bounding boxes of roughly the same size as the target at the predicted location; note that MFR does not explicitly penalize the use of large bounding boxes.

Two commonly used criteria are precision and recall scores for the tracker detected/estimated (single) target locations [50]. Precision (related to track purity) is defined as the ratio of the number of correctly tracked frames, $|TP|$, to total number of tracked frames or track length,

$$\text{Precision} = \frac{\#\ correct\ frames}{\#\ tracked\ frames} = \frac{|TP|}{|TP| + |FP|} \quad (8)$$

where number of correct frames are those in which target locations are within a set threshold distance from the ground-truth (*i.e.* 20 pixel radius ribbon). Recall (related to target purity) is the ratio of number of correctly tracked frames to number of ground-truth frames for the target defined as,

$$\text{Recall} = \frac{\#\ correct\ frames}{\#\ total\ GT\ frames} = \frac{|TP|}{|TP| + |FN|} \approx 1 - MFR. \quad (9)$$

The equality is approximate since MFR uses a bounding box overlap criteria whereas precision and recall used a distance from ground-truth centroid metric. The final performance metric used for evaluating trackers is the detected position errors defined as the distance between the estimated object position and the ground-truth centroid. *Track-based* completeness, fragmentation, mean track length, id switches, gaps and other measures of multi-target tracking performance are necessary for a more thorough evaluation of tracking performance [13].

Some LoFT (v1.3) modules (in Figure 1) were turned off for the experiments including binary classifier, background subtraction, and MHT in order to focus on evaluating the appearance update performance. LoFT performance was compared to several state-of-the-art trackers. Table II summarizes the MFR scores of eight trackers on CLIF data. Table III shows overall precision-recall scores for five of the trackers on the 14 CLIF sequences; we used author provided source code for Nearest-Neighbor (NN) [20], L1-BPR Sparse Tracker [25], Multiple Instance Learning (MILTrack) [21], and P-N Tracker [22]. We did some limited parameter tuning for optimizing each tracker for both CLIF and FMV separately. Figure 10 shows position errors three sample CLIF sequences where LoFT does particularly well. These comparisons show that our LoFT tracker outperforms all other trackers on this CLIF dataset. According to the MFR scores, MILTrack and L1-BPR Sparse trackers produced results comparable to LoFT for some of the sequences, however, the lack of a termination module causes their precision scores to drop significantly in Table III. The P-N tracker has very good performance on FMV, but the search method involves scanning the entire image and thus testing on

Method	Precision	Recall
L1-BPR [25]	0.185	0.185
MILTrack [21]	0.271	<u>0.271</u>
P-N [22]	<u>0.373</u>	0.172
NN [20]	0.088	0.082
LoFT	**0.603**	**0.405**

TABLE III
OVERALL PRECISION - RECALL SCORES ACROSS 14 CLIF SEQUENCES. SECOND BEST PERFORMANCE UNDERLINED.

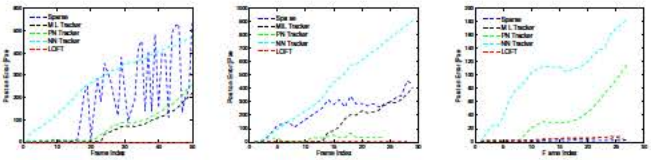

Fig. 10. Position error over the entire sequence in pixels versus frame index for five of the trackers on three selected CLIF sequences (C1_4_6, C4_4_1 and C5_3_7) for which LoFT has a high accuracy.

WAMI posed severe memory constraints. P-N tracker has the second highest precision on the CLIF data. The NN tracker had the worst results on CLIF WAMI likely due to the need to tune the SIFT features. Figure 11 shows some visual trajectories of tracking results where LoFT does well. Two sequences where LOFT did not do well, are C3_3_4 which is challenging for all of the trackers, and C4_1_0 which has many nearby spatial and temporal distractors while turning; see Figure 8 for the visual appearance of these targets and environments.

Since most published trackers are designed for standard FMV sequences, we also evaluated LoFT on several popular benchmark videos with very different scene content and characteristics compared to WAMI. Table IV shows the mean distance error to ground-truth for eight published trackers including LoFT, on four standard FMV sequences across all frames of each sequence. The PROST, AdaBoost and FragTrack results are from Gu *et al.* [20]. Figure 12 shows

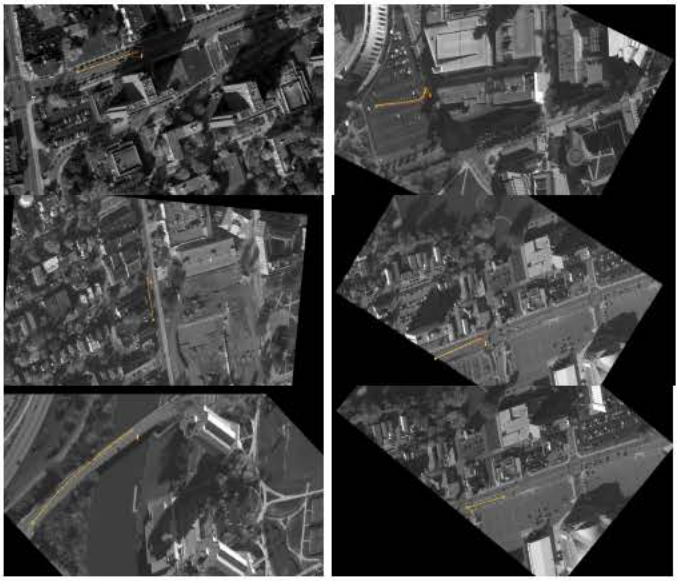

Fig. 11. LoFT results (red tracks) for six sequences. Top: C0_3_0, C1_4_6, Middle: C2_4_1, C4_4_1, Bottom: C1_2_0, C4_4_4, showing enhanced images with ground-truth tracks in yellow. LoFT outperforms other trackers in this set of sequences.

CLIF Seq.	MIL [21]	MS [51]	CPF [52]	HPF [53]	L1-BPR [11]	NN [20]	PN [22]	LoFT
C0_3_0	0.860	0.980	0.940	0.980	**0.760**	0.920	0.940	**0.760**
C1_2_0	0.852	0.963	0.9636	0.963	0.630	0.962	0.962	**0.074**
C1_4_0	0.680	0.780	0.740	0.760	**0.620**	1.000	0.700	0.820
C1_4_6	0.360	0.940	0.800	0.880	0.360	0.980	0.560	**0.040**
C2_4_1	0.900	0.980	0.980	0.980	0.920	0.980	0.980	**0.440**
C3_3_4	0.963	0.963	0.963	0.963	**0.704**	0.962	0.962	0.889
C4_1_0	**0.389**	0.889	0.833	0.889	**0.389**	0.888	0.944	1.000
C4_3_0	**0.650**	0.950	0.950	0.800	0.750	0.947	—	**0.100**
C4_4_1	0.533	0.967	0.900	0.900	**0.033**	0.931	0.758	**0.035**
C4_4_4	**0.000**	0.923	0.385	0.307	**0.000**	0.076	0.923	**0.000**
C5_1_4	**0.667**	0.958	0.875	0.833	**0.667**	0.958	0.958	0.792
C5_2_0	0.918	0.979	0.959	0.979	0.979	0.979	—	0.918
C5_3_7	**0.000**	0.963	0.148	**0.000**	**0.000**	0.8516	0.259	**0.037**
C5_4_1	**0.000**	0.952	0.810	0.905	0.958	0.523	0.809	**0.000**
Mean	0.555	0.942	0.803	0.796	0.555	0.854	0.813	**0.422**
OverAll	0.627	0.940	0.833	0.837	0.611	0.909	0.680	**0.473**

TABLE II

MISSING FRAME RATE (MFR) ON CLIF WAMI (LOWER IS BETTER). RESULTS FOR MULTIPLE INSTANCE LEARNING TRACKER (MIL), MEAN SHIFT TRACKER (MS), COVARIANCE BASED PARTICLE FILTER (CPF) TRACKER, HISTOGRAM-BASED PARTICLE FILTER (HPF) TRACKER AND ℓ_1-BOUNDED PARTICLE RESAMPLING (L1-BPR OR SPARSE) TRACKER ARE FROM LING *et al.* [11]. MEAN INDICATES AVERAGE OF SEQUENCE MFRS (SHORTER TRACKS HAVE HIGHER INFLUENCE) WHILE OVERALL IS AN ENSEMBLE AVERAGE AS IN [11].

Sequence	PROST [54]	AdaBoost [55]	FragTrack [56]	MILTrack [21]	NN [20]	L1-BPR [24]	PN [22]	LoFT
Girl	19.00	43.30	26.50	31.60	<u>18.00</u>	67.84	28.88	**13.86**
David	<u>15.30</u>	51.00	46.00	15.60	15.60	63.12	**10.38**	40.60
Faceocc	<u>7.00</u>	49.00	**6.50**	18.40	10.00	20.78	13.99	10.79
Faceocc2	17.20	19.60	45.10	14.30	**12.90**	73.27	19.14	<u>13.25</u>

TABLE IV

MEAN POSITION ERRORS ON STANDARD FULL MOTION VIDEOS WITH PROST, ADABOOST, FRAGTRACK, MILTRACK AND NN RESULTS FROM GU *et al.* [20]. BEST RESULTS AND SECOND BEST RESULTS ARE SHOWN IN BOLD AND UNDERLINED RESPECTIVELY.

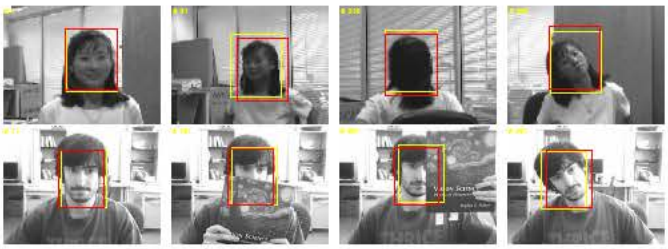

Fig. 12. Tracking results showing sample frames from 'girl' and 'faceocc2' sequences showing bounding boxes for ground-truth (yellow) and LoFT (red).

sample frames from LoFT tracking results compared to GT for 'girl' and 'faceocc2' sequences. Instead of tight initial bounding boxes we used the actual GT bounding box on the appropriate start frame in each FMV sequence. Based on the mean distance errors, the LoFT system is comparable to the other trackers on these four representative FMV sequences. LoFT also produced better results for these two videos.

VI. CONCLUSIONS

We described our Likelihood of Features Tracking (LoFT) system developed to track vehicles of interest in challenging low frame rate aerial WAMI, within a single target tracking context. LoFT uses an adaptive set of feature descriptors with posterior fusion modeled as recognition-based track-before-detect, a novel appearance and pose estimation algorithm, coupled with a track management module to achieve much better performance compared to other state-of-the-art trackers including L1-BPR, a sparse representation-based tracking approach also adapted for WAMI, and learning-based tracking algorithms like MILTrack and P-N Tracker. On the CLIF dataset LoFT improves on the best previous results by 13.8% (L1-BPR) and 15.4% (MILTrack) using the MFR metric. In terms of precision and recall LoFT was 23.0% and 13.4% higher respectively, compared to the second best trackers. On FMV data LoFT performs quite competitively with the best trackers in the literature and parameters were not customized or tuned specifically for FMV in these preliminary tests. The versatility of our approach for a range of tracking tasks was demonstrated by LoFT's competitive performance on both WAMI and FMV sequences with very different foreground-background characteristics, camera geometry and framerates. LoFT is not restricted to single target tracking and can be readily extended to multi-target tracking using multiple trackers running in parallel in a supervised manner, or given a collection of detections or moving target indicators, for unsupervised automatic tracking.

ACKNOWLEDGMENTS

The fourteen CLIF vehicle sequences with registration homographies and ground-truth bounding boxes used in the experiments were kindly provided by Haibin Ling at Temple University. Rui Wang and Raphael Viguier assisted with testing several of the trackers and Anoop Haridas improved KOLAM for visualizing and comparing trajectories. This research was partially supported by U.S. Air Force Research Laboratory (AFRL) under agreement AFRL FA8750-11-C-0091. Approved for public release (case 88ABW-2012-1013). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies of AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey.," *ACM Comput. Surv.*, vol. 38, 2006.

[2] R. Kumar and *et al.*, "Aerial video surveillance and exploitation," *Proc. of the IEEE*, vol. 89, no. 10, pp. 1518–1539, 2001.

[3] K. Palaniappan, R. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks: Research Challenges and Future Directions*, B. Banhu and et al., Eds., chapter 24, pp. 349–371. Springer, 2011.

[4] K. Palaniappan and *et al.*, "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," in *13th Conf. on Information Fusion*, 2010, pp. 1–8.

[5] A. Haridas, R. Pelapur, J. Fraser, F. Bunyak, and K. Palaniappan, "Visualization of automated and manual trajectories in wide-area motion imagery," in *Int. Conf. Information Visualisation*, 2011, pp. 288–293.

[6] J. Fraser, A. Haridas, G. Seetharaman, R. Rao, and K. Palaniappan, "KOLAM: An extensible cross-platform architecture for visualization and tracking in wide-area motion imagery," in *Proc. SPIE Conf. Geospatial InfoFusion II*, 2012, vol. 8396.

[7] E.P. Blasch, P.B. Deignan, S.L. Dockstader, M. Pellechia, K. Palaniappan, and G. Seetharaman, "Contemporary concerns in geographical/geospatial information systems (GIS) processing," in *Proc. IEEE National Aerospace and Electronics Conf.*, 2011, pp. 183–190.

[8] R. Porter, A. M. Fraser, and D. Hush, "Wide-area motion imagery," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 56–65, 2010.

[9] N.P. Cuntoor, A. Basharat, A.G.A. Perera, and A. Hoogs, "Track initialization in low frame rate and low resolution videos," in *Int. Conf. Pattern Recognition*. IEEE, 2010, pp. 3640–3644.

[10] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," *Proc. European Conf. Computer Vision*, pp. 186–199, 2010.

[11] H. Ling and *et al.*, "Evaluation of visual tracking in extremely low frame rate wide area motion imagery," in *14th Int. Conf. on Information Fusion*, 2011, pp. 1–8.

[12] J. Prokaj and G. Medioni, "Using 3D scene structure to improve tracking," in *Proc. IEEE CVPR*, 2011.

[13] E.K. Kao, M. P. Daggett, and M. B. Hurley, "An information theoretic approach for tracker performance evaluation," in *Proc. Int. Conf. Computer Vision*, 2009, pp. 1523–1529.

[14] E. Pollard, A. Plyer, B. Pannetier, F. Champagnat, and G. Le Besnerais, "GM-PHD filters for multi-object tracking in uncalibrated aerial videos," in *12th Int. Conf. Information Fusion*, 2009, pp. 1171–1178.

[15] R. Porter, C. Ruggiero, and J.D. Morrison, "A framework for activity detection in wide-area motion imagery," *Proc. SPIE Defense, Security and Sensing*, p. 73410O, 2009.

[16] G.D. Hager and P.N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. PAMI*, vol. 20, no. 10, pp. 1025–1039, 1998.

[17] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Trans. PAMI*, vol. 23, no. 6, pp. 681 –685, 2001.

[18] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[19] R Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. PAMI*, vol. 27, pp. 1631–1643, 2005.

[20] S. Gu, Y. Zheng, and C. Tomasi, "Efficient visual object tracking with online nearest neighbor classifier," in *Proc. Asian Conf. Computer Vision*, 2010, pp. 271–282.

[21] B. Babenko, M. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. PAMI*, vol. 33, no. 8, pp. 1619–1631, 2011.

[22] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," *Proc. IEEE CVPR*, pp. 49–56, 2010.

[23] I. Ersoy, K. Palaniappan, and G. Seetharaman, "Visual tracking with robust target localization," in *IEEE Int. Conf. Image Processing*, 2012.

[24] Xue Mei, Haibin Ling, Yi Wu, Erik Blasch, and Li Bai, "Minimum error bounded efficient l_1 tracker with occlusion detection," in *IEEE Proc. CVPR*, 2011.

[25] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," in *IEEE Trans. PAMI*, 2011, pp. 2259–2272.

[26] A. Hafiane, G. Seetharaman, K. Palaniappan, and B. Zavidovique, "Rotationally invariant hashing of median patterns for texture classification," *Lecture Notes in Computer Science*, vol. 5112, pp. 619–629, 2008.

[27] Y. Boers, F. Ehlers, W. Koch, T. Luginbuhl, L.D. Stone, and R.L. Streit, "Track before detect algorithms," *EURASIP Journal on Advances in Signal Processing*, 2008.

[28] J. Kittler, M. Hatef, R.P.W. Duin, and Matas J, "On combining classifiers," *IEEE Trans. PAMI*, vol. 20(3), pp. 226–239, 1998.

[29] Z. Yin, F. Porikli, and R. Collins, "Likelihood map fusion for visual object tracking," in *IEEE Workshop Appl. Comput. Vis.*, 2008, pp. 1–7.

[30] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. BMVC*, 2006, vol. 1, pp. 47–56.

[31] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. PAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[32] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization," *Proc. European Conf. Computer Vision*, pp. 624–637, 2010.

[33] R. Pelapur, K. Palaniappan, F. Bunyak, and G. Seetharaman, "Vehicle orientation estimation using radon transform-based voting in aerial imagery," in *SPIE conf. on Geospatial Infofusion*, 2012, p. accepted.

[34] P. Bellens, K. Palaniappan, R. M. Badia, G. Seetharaman, and J. Labarta, "Parallel implementation of the integral histogram," *Lecture Notes in Computer Science (ACIVS)*, vol. 6915, pp. 586–598, 2011.

[35] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cognitive science*, vol. 11, no. 1, pp. 23–63, 1987.

[36] F. Bunyak and K. Palaniappan, "Efficient segmentation using feature-based graph partitioning active contours," in *12th IEEE Int. Conf. Computer Vision*, Kyoto, Japan, 2009, pp. 873–880.

[37] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *J. Multimedia*, vol. 2, no. 4, pp. 20–33, August 2007.

[38] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Radon transform orientation estimation for rotation invariant texture analysis," *IEEE Trans. PAMI*, vol. 27, no. 6, pp. 1004–1008, 2005.

[39] R. Pelapur, K. Palaniappan, and G. Seetharaman, "Robust orientation and appearance adaptation for wide-area large format video object tracking," in *IEEE Conf. Adv. Video Signal-based Surveillance*, 2012.

[40] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, Wiley-Interscience, 2001.

[41] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep., Univ. of North Carolina, Chapel Hill, 1995.

[42] Air Force Research Laboratory, "Columbus Large Image Format (CLIF) dataset over Ohio State University," 2007.

[43] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[44] O. Mendoza-Schrock, J.A. Patrick, and E.P. Blasch, "Video image registration evaluation for a layered sensing environment," in *Proc. IEEE National Aerospace Electronics Conf. (NAECON)*, 2009, pp. 223–230.

[45] G. Seetharaman, G. Gasperas, and K. Palaniappan, "A piecewise affine model for image registration in 3-D motion analysis," in *IEEE Int. Conf. Image Processing*, 2000, pp. 561–564.

[46] A. Hafiane, K. Palaniappan, and G. Seetharaman, "UAV-video registration using block-based features," in *IEEE Int. Geoscience and Remote Sensing Symposium*, 2008, vol. II, pp. 1104–1107.

[47] K. Palaniappan and J.B. Fraser, "Multiresolution tiling for interactive viewing of large datasets," in *17th Int. AMS Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*. American Meteorological Society, 2001, pp. 338–342.

[48] C. Liu, W.T. Freeman, E.H. Adelson, and Y. Weiss, "Human-assisted motion annotation," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[49] S. Candemir, K. Palaniappan, F. Bunyak, and G. Seetharaman, "Feature fusion using ranking for object tracking in aerial imagery," in *Proc. SPIE Conf. Geospatial InfoFusion II*, 2012, vol. 8396.

[50] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*, Wiley, 2011.

[51] Chen D and J. Yang, "Robust object tracking via online dynamics spatial bias appearance models," *IEEE Trans. PAMI*, vol. 29, no. 12, pp. 2157–2169, 2007.

[52] Y. Wu, B. Wu, J. Liu, and H.Q. Lu, "Probabilistic tracking on riemannian manifolds," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2008.

[53] P. Perez and *et al.*, "Color-based probabilistic tracking," in *Proc. European Conf. Computer Vision*, 2002.

[54] J. Santner and *et al.*, "PROST: parallel robust online simple tracking," in *Proc. IEEE CVPR*, 2010.

[55] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE CVPR*, 2006, pp. 260–267.

[56] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE CVPR*, 2006, pp. 798–805.